SUMMARY=ASYMPTOTIC STATISTICS-13-RANK

GEN RYU

Consider that we have N samples $(X_1, \dots, X_N)$.

1. REFERENCES

(1) Asymptotic statistics メインテキスト
(2) Theory of rank tests サブテキスト (明らかに上のテキストより詳しい)

2. WORDS

(1) tie with と同じである。
(2) remedy 修正する
(3) subsequently その次に

3. DEFINITIONS

3.1. ordered statistics.

$$X_{N(1)} \leq \cdots \leq X_{N(N)},$$

which makes a seq in the increasing order is an asymptotic case. If the asymptotic property is not considered, then we may write as follows in the same meaning:

$$X^{(1)} \leq \cdots \leq X^{(N)}.$$

3.2. rank. R will stand for the vector of ranks $(R_1, \dots, R_N)$. r and $(r_1, \dots, r_N)$ will be the realization of R, respectively. in the asymptotic case we use

$$R_{Ni}.$$

It is the position number of the i-th sample in N samples. The property is

$$X_i = X_{N(R_{Ni})}.$$

note1. If X_i is tied with some other observations, then we can not define the rank uniquely. In this case, we have two ways to solve this problem as follows:

(1) Let $X_i = X_{N(j)}$ for all i, where X_i have the same value, and j is the average of all ranks that the sample takes;
(2) Let $R_{Ni} = \sum_{j=1}^{N} 1_{\{X_j \leq X_i\}}$. (uprank)

However, we will assume the distribution from which the sample is is continuous. It means that the case will be a null set.

note2. the pair $(X^{(i)}, R)$ is a sufficient statistic for any system of distributions determined by densities.

Date: April 8, 2012.

3.3. **linear rank statistic.**

$$\sum_{i=1}^{N} a_N(i, R_{Ni})$$

for a given $(N \times N)$ matrix $(a_N(i,j))$. This is the sum of the elements of the matrix $(a_N(i,j))$.

Example 1. *Let* $X = (2,3,1)$. *Then*

$$\begin{aligned}(1, R_{N1}) &= (1,2)\\ (2, R_{N2}) &= (2,3)\\ (3, R_{N3}) &= (3,1).\end{aligned}$$

Thus, the position of the matrix can be shown as

$$\begin{pmatrix} a_{11} & a_{12} & \bigcirc \\ \bigcirc & a_{22} & a_{23} \\ a_{31} & \bigcirc & a_{33} \end{pmatrix}.$$

3.4. **simple linear rank statistics.**

$$\sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}};$$

This is a form of the sum of the elements of the matrix multiplied by some coefficients.

3.5. **coefficients.**

$$(c_{N1}, \ldots, c_{NN});$$

3.6. **scores.**

$$(a_{N1}, \ldots, a_{NN}).$$

3.7. **i-th smallest coordinate.** [not AS]

$$o_i(x),$$

and obviously $x^{(i)} = o_i(x)$.

3.8. **the system where the distribution of $(X_1, \ldots, X_N)$ is symmetric and determined by a density.**

$$p(x_{r_1}, \ldots, x_{r_N}) = p(x_1, \ldots, x_N), \quad r \in R,$$

if and only if $p \in H_*$.

3.9. **the system where the observation is iid.**

$$p = \prod_{i=1}^{N} f(x_i),$$

where f(x) may be an arbitrary one-dimensional density if and only if $p \in H_0$. *note.* $H_0 \subset H_*$.

3.10. incomplete Beta function ratio I_z(a,b).

$$F(x) = I_x(a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\,dt}{B(a,b)}, \quad 0 \le x \le 1; a,b > 0.$$

The mean of Beta distribution is $\frac{a}{a+b}$, the mode is $\frac{a-1}{a+b-2}$, the variance is $\frac{ab}{(a+b)^2(a+b+1)}$, the coefficient of variation is $\sqrt{\frac{b}{a(a+b+1)}}$, and the skewness is $\frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$

4. LEMMAS

Lemma 4.1. *If X is governed by the density q, then $X^{(\cdot)}$ is governed by the density*

$$\bar{q}(x^{(1)}, \ldots, x^{(N)}) \triangleq \sum_{r \in R} q(x^{(r_1)}, \ldots, x^{(r_N)}), \quad x^{(\cdot)} \in \boldsymbol{X}^{(\cdot)}.$$

Moreover,

$$Q(R = r | X^{(\cdot)} = x^{(\cdot)}) = \frac{q(x^{(r_1)}, \ldots, x^{(r_N)})}{\bar{q}(x^{(1)}, \ldots, x^{(N)})}, \quad r \in R, x^{(\cdot)} \in \boldsymbol{X}^{(\cdot)},$$

holds with Q being the probability distribution corresponding to q.

Proof. For any $A \in \mathcal{A}^{(\cdot)}$, it holds that

$$\begin{aligned}
\int \cdots \int_{X^{(\cdot)} \in A} q(x_1, \ldots, x_N) dx_1 \ldots dx_N &= \sum_{r \in R} \int \cdots \int_{X^{(\cdot)} \in A, R = r} q(x_1, \ldots, x_N) dx_1 \ldots dx_N \\
&= \sum_{r \in R} \int \cdots \int_A q(x^{(r_1)}, \ldots, x^{(r_N)}) dx^{(1)} \ldots dx^{(N)}
\end{aligned}$$

Note that the Jacobian is 1 in this case. □

note. $\bar{q}$ do not have to be equal to each other. See the example following.

Example 2. *Let $\Omega = (1,2,3)$ and the probability on it is defined as*

$$(p_1, p_2, p_3) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}).$$

After taking one sample from Ω, we will have the sample set like one of the following three cases:

$$\begin{aligned}
\Omega' &= (1,3), & (p_1, p_3) &= (\frac{1}{3}, \frac{2}{3}); \\
\Omega' &= (2,3), & (p_2, p_3) &= (\frac{1}{3}, \frac{2}{3}); \\
\Omega' &= (1,2), & (p_1, p_2) &= (\frac{1}{2}, \frac{1}{2}).
\end{aligned}$$

Then the probability for $(x^{(1)}, x^{(2)})$ *will be* $\frac{5}{12}$*, and the probability for* $(x^{(2)}, x^{(1)})$ *will be* $\frac{7}{12}$*. Furthermore,*

$$\begin{aligned}\bar{q}(1,2) &= \frac{1}{6};\\ \bar{q}(2,3) &= \frac{5}{12};\\ \bar{q}(1,3) &= \frac{5}{12}.\end{aligned}$$

This example is a special case, and what we will think next is the property on the system 3.8 and 3.9.

Lemma 4.2. *Let* $X_1, \ldots, X_N$ *be a random sample from a continuous distribution function* F *with density* f*. Then*

(1) *the vectors* $X_{N()}$ *and* R_N *are independent;*
(2) *the vector* $X_{N()}$ *has density* $N!\prod_{i=1}^{N} f(x_i)$ *on the set* $x_1 < \cdots < x_N$*;*
(3) *the variable* $X_{N(i)}$ *has density* $N\binom{N-1}{i-1}F(x)^{i-1}(1-F(x))^{N-i}f(x)$*; for* F *the uniform distribution on* $[0,1]$*, it has mean* $i/(N+1)$ *and variance* $i(N-i+1)/((N+1)^2(N+2))$*;*
(4) *the vector* R_N *is uniformly distributed on the set of all* $N!$ *permutations of* $1, 2, \ldots, N$*;*
(5) *for any statistic* T *and permutation* $r = (r_1, \ldots, r_N)$ *of* $1, 2, \ldots, N$*,*
$$E(T(X_1, \ldots, X_N)|R_N = r) = ET(X_{N(r_1)}, \ldots, X_{N(r_n)});$$
(6) *for any simple linear rank statistic* $T = \sum_{i=1}^{N} c_{Ni}a_{N,R_{Ni}}$*,*
$$ET = N\bar{c}_N\bar{a}_N; \quad \mathrm{Var}T = \frac{1}{N-1}\sum_{i=1}^{N}(c_{Ni}-\bar{c}_N)^2\sum_{i=1}^{N}(a_{Ni}-\bar{a}_N)^2.$$

Proof.

(1) It is obvious from Lemma 1.
(2) The same as the above.
(3)
(4)
(5) Just change the rotation of the random variables, then we can see it by the virtue of the independence between $X_{N()}$ and R_N.

□

note. Even we suppose the density is identical, (1), (2), (5) of the lemma hold even in the case of 3.8.

Corollary 4.3. *As the same condition, the variable* $X_{N(i)}$ *has density*

$$F_{N(i)}(x) = I_{F(x)}(i, N-i+1) = \frac{N!}{(i-1)!(N-i)!}\int_0^{F(x)} u^{i-1}(1-u)^{N-i}\,du.$$

5. A NECESSARY CONDITION FOR RANK STATISTICS ASYMPTOTICALLY NORMAL

The scores are generated through a given function $\phi : [0,1] \to \mathbb{R}$ in one of two ways. Either

$$a_{Ni} = E\phi(U_{N(i)}), \tag{5.1}$$

where $U_{N(1)}, \ldots, U_{N(N)}$ are the order statistics of a sample of size N from the uniform distribution on $[0,1]$; or

$$a_{Ni} = \phi(\frac{i}{N+1}). \tag{5.2}$$

For well-behaved functions ϕ, these two definitions are closely related and almost identical, since $EU_{N(i)} = \frac{i}{N+1}$.

note. Scores of the first type correspond to the locally most powerful rank tests; scores of the second type are attractive in view of their simplicity.

6. Property of rank statistics

Theorem 6.1. *Let R_N be the rank vector of an i.i.d. sample $X_1, \ldots, X_N$ from the continuous distribution function F. Let the scores a_N be generated according to (**??**) for a measurable function ϕ that is not constant almost everywhere, and satisfies $\int_0^1 \phi^2(u)\,du < \infty$. Define the variables*

$$T_N = \sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}}, \quad \tilde{T}_N = N\bar{c}_N \bar{a}_N + \sum_{i=1}^{N} (c_{Ni} - \bar{c}_N)\phi(F(X_i)).$$

Then the sequences T_N and $\tilde{T}_N$ are asymptotically equivalent in the sense that

- $$ET_N = E\tilde{T}_N$$
- $$\frac{\mathrm{Var}(T_N - \tilde{T}_N)}{\mathrm{Var}T_N} \to 0$$

.

*The same is true if the scores are generated according to (**??**) for a function ϕ that is continuous and almost everywhere, is nonconstant, and satisfies*

$$\frac{1}{N}\sum_{i=1}^{N} \phi^2(\frac{i}{N+1}) \to \int_0^1 \phi^2(u)\,du < \infty.$$

Theorem 6.2 (Lindeberg-Feller central limit theorem)**.** *For each n let $Y_{n,1}, \ldots, Y_{n,k_n}$ be independent random vectors with finite variances such that*

$$\sum_{i=1}^{k_n} E||Y_{n,i}||^2 1\{||Y_{n,i} > \epsilon||\} \to 0, \quad \textit{for every } \epsilon > 0,$$
$$\sum_{i=1}^{k_n} \mathrm{Cov} Y_{n,i} \to \Sigma.$$

Then

$$\sum_{i=1}^{k_n}(Y_{n,i} - EY_{n,i}) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Note that

$$\frac{\max_{1\le i\le N(c_{Ni}-\bar{c}_N)^2}}{\sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2} \to 0. \tag{6.1}$$

Corollary 6.3. *If the vector of coefficients c_N satisfies (**??**), and the scores are generated according to (**??**) for a measurable, nonconstant, square-integrable function ϕ, then the sequence of standardized rank statistics*

$$\frac{(T_N - ET_N)}{\mathrm{sd}T_N} \xrightarrow{d} \mathcal{N}(0, 1).$$

*The same is true if the scores are generated by (**??**) for a function ϕ that is continuous almost everywhere, is nonconstant, and satisfies*

$$\frac{1}{N}\sum_{i=1}^N \phi^2(\frac{i}{N+1}) \to \int_0^1 \phi^2(u)\,du.$$

7. Signed Rank Statistics

Lemma 7.1. *Let $X_1, \dots, X_N$ be a random sample from a continuous distribution that is symmetric about 0. Then*

1. *the vectors $(|X|, R_N^+)$ and $\mathrm{sign}_N(X)$ are independent;*
2. *the vector R_N^+ is uniformly distributed over $\{1, \dots, N\}$;*
3. *the vector $\mathrm{sign}_N(X)$ is uniformly distributed over $\{-1, 1\}^N$;*
4. *for any signed rank statistic, $\mathrm{Var}\sum_{i=1}^N a_{N,R_{Ni}^+}\mathrm{sign}(X_i) = \sum_{i=1}^N a_{Ni}^2$.*

Theorem 7.2. *Let $X_1, \dots, X_N$ be a random sample from a continuous distribution that is symmetric about 0. Let the scores a_N be generated according to (**??**) for a measurable function ϕ such that $\int_0^1 \phi^2(u)\,du < \infty$. For F^+ the distribution function of $|X_1|$, define*

$$T_N = \sum_{i=1}^N a_{N,R_{Ni}^+}\mathrm{sign}(X_i), \quad \tilde{T}_N = \sum_{i=1}^N \phi(F^+(|X_i|))\mathrm{sign}(X_i).$$

Then the sequences T_N and $\tilde{T}_N$ are asymptotically equivalent in the sense that $\frac{1}{N}\mathrm{Var}(T_N - \tilde{T}_N) \to 0$. Consequently, the sequence

$$N^{-1/2}T_N \xrightarrow{d} \mathcal{N}(0, \int_0^1 \phi^2(u)\,du).$$

The same is true if the scores are generated according to (**??**) *for a function* ϕ *that is continuous almost everywhere and satisfies*

$$\frac{1}{N}\sum_{i=1}^{N}\phi^2(\frac{i}{N+1}) \to \int_0^1 \phi^2(u)\,du.$$

8. Rank statistics under alternatives

8.1. **smooth score-generating functions.** Let $\bar{F}_N$ be the average of $F_1, \dots, F_N$ and let $\bar{F}_N^c$ be the weighted sum $N^{-1}\sum_{i=1}^{N} c_{Ni}F_i$, and define

$$T_N = \sum_{i=1}^{N} c_{Ni}\phi(\frac{R_{Ni}}{N+1}), \quad \hat{T}_N = \sum_{i=1}^{N}\left[c_{Ni}\phi(\bar{F}_N(X_i)) + \int_{X_i}^{\infty}\phi'(\bar{F}_N(x))\,d\bar{F}_N^c(x)\right].$$

Here, the variables $\hat{T}_N$ are the Hájek projections of approximations to the variables T_N, up to centering at mean 0. (あくまでも近似で、ちゃんとした projection については、結果はよりよいが、形はもっと煩雑である。)

Lemma 8.1. *If* $\phi : [0,1] \mapsto \mathbb{R}$ *is twice continuously differentiable, then there exists a universal constant* K *such that*

$$\mathrm{Var}(T_N - \hat{T}_N) \le K\frac{1}{N}\sum_{i=1}^{N}(c_{Ni} - \bar{c}_N)^2(||\phi'||_\infty^2 + ||\phi''||_\infty^2).$$

note. As a consequence of the lemma, the sequences

$$\frac{T_N - ET_N}{\mathrm{sd}T_N} \quad \text{and} \quad \frac{\hat{T}_N - E\hat{T}_N}{\mathrm{sd}\hat{T}_N}$$

have the same limiting distribution (if any) if

$$\frac{\sum_{i=1}^{N}(c_{Ni} - \bar{c}_N)^2}{N\mathrm{Var}\hat{T}_N} \to 0.$$

Lemma 8.2 (Variance Inequality)**.** *For nondecreasing coefficients* $a_{N1} \le \cdots \le a_{NN}$ *and arbitrary scores* $c_{N1}, \dots, c_{NN}$*,*

$$\mathrm{Var}\sum_{i=1}^{N} c_{Ni}a_{N,R_{Ni}} \le 21 \max_{1\le i\le N}(c_{Ni} - \bar{c}_N)^2\sum_{i=1}^{N}(a_{Ni} - \bar{a}_N)^2.$$

Theorem 8.3 (Rank central limit theorem)**.** *Let* $T_N = \sum c_{Ni}a_{N,R_{Ni}}$ *be the simple linear rank statistic with coefficients and scores such that*

$$\max_{1\le i\le N}\frac{|a_{Ni} - \bar{a}_N|}{\sum_{i=1}^{n}(a_{Ni} - \bar{a}_N)^2} \to 0;$$

$$\max_{1\le i\le N}\frac{|c_{Ni} - \bar{c}_N|}{\sum_{i=1}^{N}(c_{Ni} - \bar{c}_N)^2} \to 0.$$

Let the rank vector R_N be uniformly distributed on the set of all $N!$ permutations of $\{$ 1, 2, $\ldots$, $N\}$. Then the sequence

$$\frac{T_N - ET_N}{\text{sd}T_N} \xrightarrow{d} \mathcal{N}(0,1),$$

if and only if, for every $\epsilon > 0$,

$$\sum_{(i,j):\sqrt{N}|a_{Ni}-\bar{a}_N||c_{Ni}-\bar{c}_N|>\epsilon A_N C_N} \frac{|a_{Ni}-\bar{a}_N|^2|c_{Ni}-\bar{c}_N|^2}{A_N^2 C_N^2} \to 0,$$

where

$$A_N^2 = \sum_{i=1}^{n}(a_{Ni}-\bar{a}_N)^2, \quad C_N^2 = \sum_{i=1}^{N}(c_{Ni}-\bar{c}_N)^2.$$

9. Main Ideas

(1) Rank test is used to test hypotheses.
(2) Generally, the null hypothesis is to assume the r.v. is iid.
(3) The virtue of using rank test is that it is distribution free.
(4) Rank test is a special example of permutation test.